HUMAN COMPUTER INTERACTION

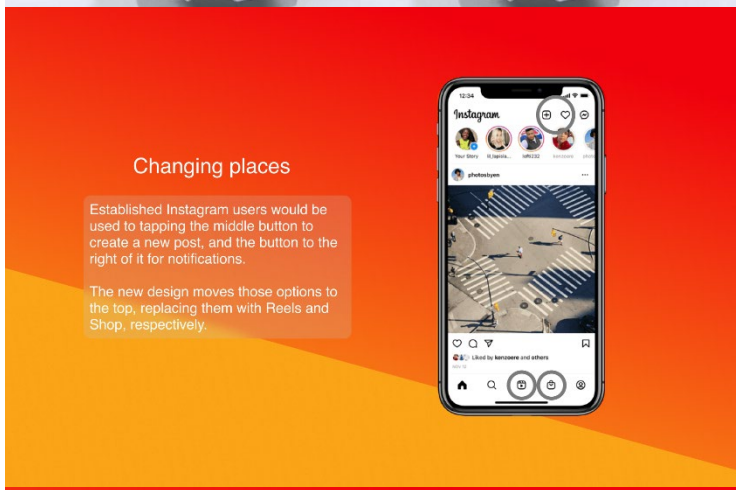
Lecture 10 Overview: Evaluation

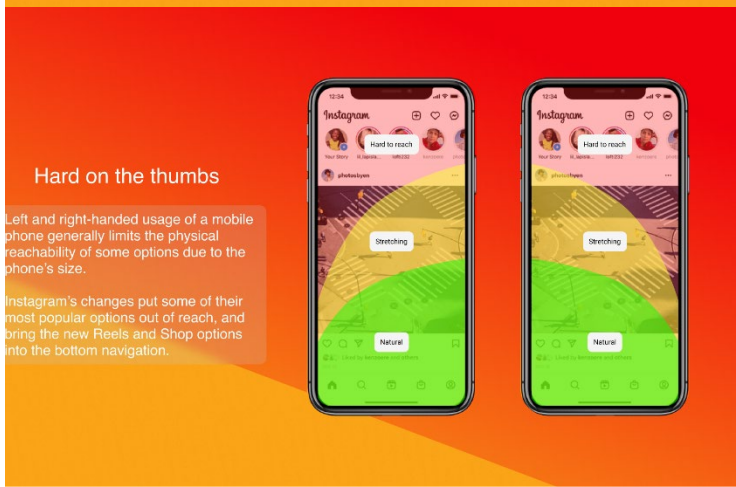
- Challenges of Evaluation
- Evaluation: Methods & Measures
  - Heuristics
  - Usability
  - Experiments
- Evaluation Results
- Evaluation Framework: DECIDE

WHY?

EVALUATION: RATIONALE

- Why evaluate? We already know what works and what doesn't, what users like and don't?
- Established products / platforms: strong user base, why evaluate?
- Examples:
  - PS 5: Inserting Disc
  - Instagram: Reels & Shopping
  - Other?





EVALUATION

- HCI is difficult to design
- Content, meaning, insight, experience central to design success, not technology: these are central competences for designers
- Designers can fail to evaluate objects objectively
- Evaluation needs to be frequent & varied
- Key Questions:
  - Why? What? Who? Where? When?

and then

  - How?

EVALUATION: CHALLENGES

- Evaluation viewed as common sense: assumptions
- Purpose of evaluation: Goals - Users or Business Need?
- Easier to build this way: human factors come first
- Testing on yourself: you are not typical
- Evaluation is carried out on wrong people - non-representative
- Lack of time to conduct evaluations or consider evaluation results
- No universal measure of usability

WHAT?

WHAT TO EVALUATE?

- Useful
- Usable: Usability
- Findable
- Credible
- Desirable
- Accessible
- Valuable

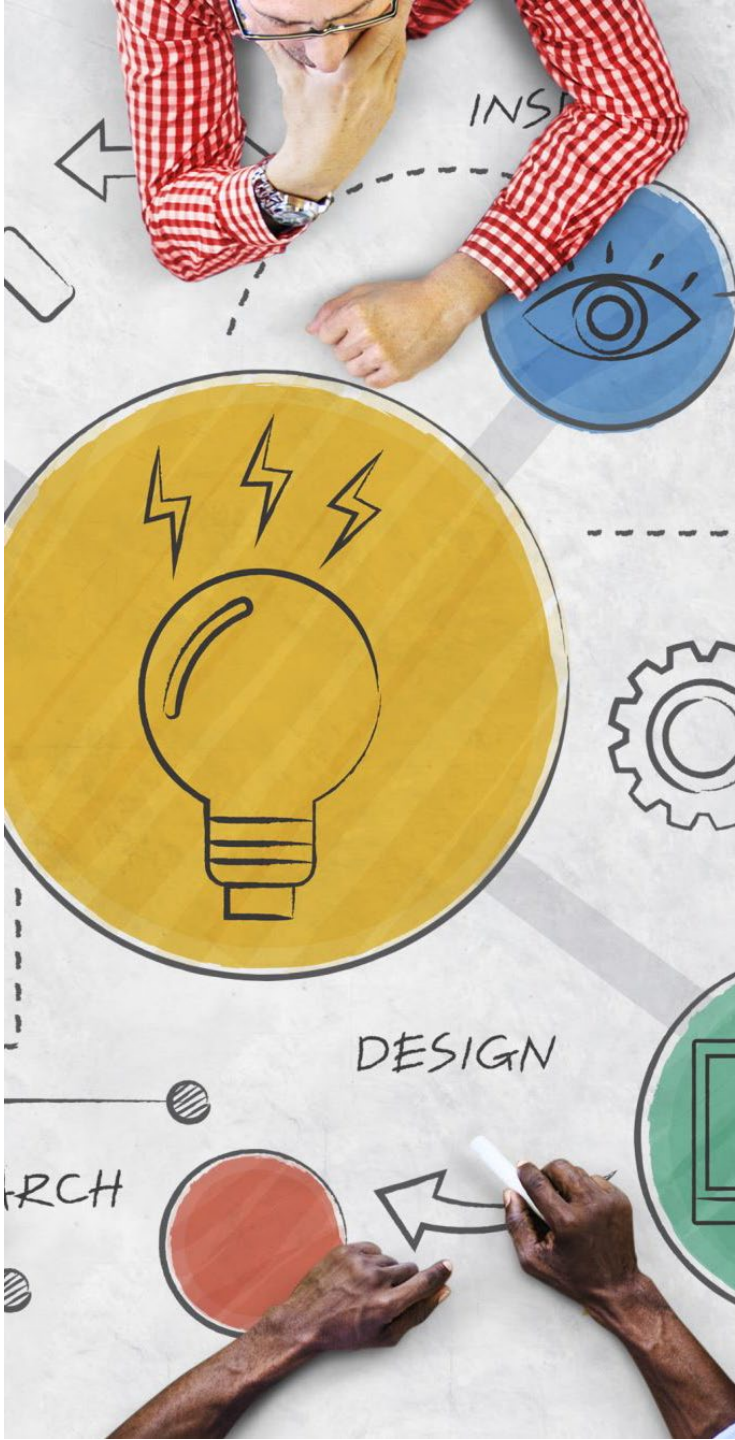
(Peter Morville)

WHAT TO EVALUATE?

- Usability: Property of System or Usage?
- Essential (homogeneous) Vs. Relational (heterogeneous)
- System or Interaction?
- User Interaction:
  - Cognitive
  - Social
  - Affective
  - Physical
  - All of the above: UX

WHAT TO EVALUATE?

- Artefact: design idea or prototype or system
- System Usability:
  - Effectiveness: Usefulness, Usability
  - Efficiency
  - User Satisfaction: UX, Used
- Data:
  - Quantitative
  - Qualitative



WHAT TO EVALUATE?

- You can't test everything!
- How do you decide what to test?
- Testing Goals: Prioritise
  - Can you change it?
  - Risk
  - Business Impact
  - User Impact

HOW?

EVALUATION METHOD CHOICE

- Design Vs Implementation
- User Vs Expert
- Laboratory Vs Field study (control Vs naturalness)
- Response: Immediate vs Delayed

EVALUATION METHOD CHOICE

- Evaluation Technique:
  - Subjective vs Objective
  - Quantitative vs Qualitative measure
  - Information Provided (detail (font) vs general (usability))
  - Immediacy of Response
  - Intrusiveness
  - Resources

EVALUATION TECHNIQUES

- Critique: Heuristics
- Query techniques (& Analytics)
- Observational techniques
- Experimental designs
- Monitoring physiological responses

EVALUATION METHODS

Method	Controlled settings	Natural settings	Without users
Observing	X	X	
Asking users	X	X	
Asking experts		X	X
Testing	X	X	
Modeling			X

EVALUATION: HEURISTIC

- General principle or rule of thumb
- An inspection-based technique for identifying usability problems in UI's
- Nielsen & Molich (1990): qualitative critique of a system using a set of relatively simple & general heuristics
- Excellent for earlier designs; high impact for low cost
- How? - Several independent evaluators critique a system – twice – to identify potential usability problems
- No evaluator finds everything; some find more than others

EVALUATION: HEURISTICS

When to Critique?

- Before user Testing: pick up small problems before users
- Before Redesigning: what works and what needs to change
- Provide Evidence: articulate problems, ammunition for redesign
- Before Release: smooth product before release

EVALUATION: HEURISTICS

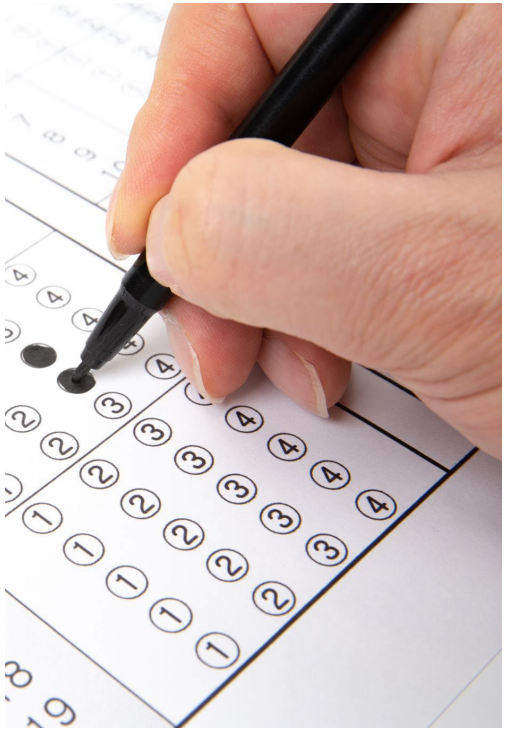
- Kritsch evaluated 10 Heuristic Frameworks; 91 heuristics (UXDesign.com)
- 6 Usability Components:

1. Learnability: referenced by 35% of all usability attributes (6.5 of 10 frameworks)
2. Efficiency: referenced by 24%
3. Satisfaction: referenced by 21%
4. Utility: referenced by 7%
5. Errors: referenced by 9%
6. Memorability: referenced by 4%

EVALUATION: HEURISTICS

- Which framework? Nielsen: holistic
- Mix and match heuristics. Add, remove, validate, adjust for specific use cases and domains.

- Use Rams to improve user **Satisfaction**
- Use the **Learnability** frameworks to focus evaluation efforts around intuitiveness (Bastien & Scapin, ISO, Kaniasty, Nielsen, Shneiderman, SUS, or Bouccher)
- Try the System Usability Scale for a quick, **quantifiable review** with a Learnability focus
- Boucher's criteria for a simplified and balanced evaluation of **Efficiency** and **Learnability**

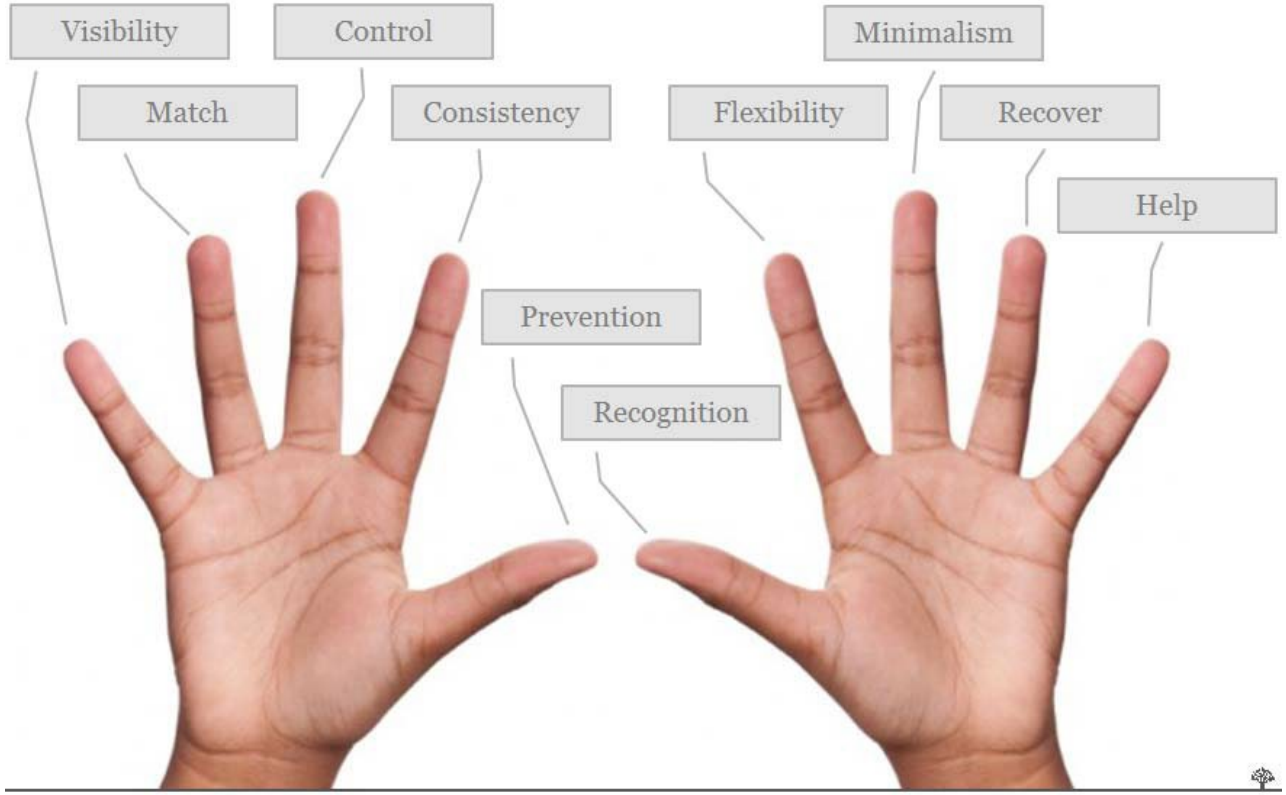


EVALUATION: HEURISTIC

Ten Heuristics (Nielsen):

1. **Visibility** of system status
2. **Match** between system and real world
3. User **control** and freedom
4. **Consistency** and standards
5. **Error prevention**
6. **Recognition** rather than recall
7. **Flexibility** and efficiency of use: shortcuts
8. **Aesthetic** and minimalist design
9. Help users recognize and **recover** from **errors**
10. **Help** and documentation

https://www.nngroup.com/articles/ten-usability-heuristics/



EVALUATION: HEURISTIC

EVALUATION: HEURISTIC

- Severity ratings: 4: catastrophic – 1: cosmetic

  4 – usability catastrophe

  3 – major usability problem

  2 – minor usability problem

  1 – cosmetic problem only

  0 – this is not a usability problem

- Number of evaluators?

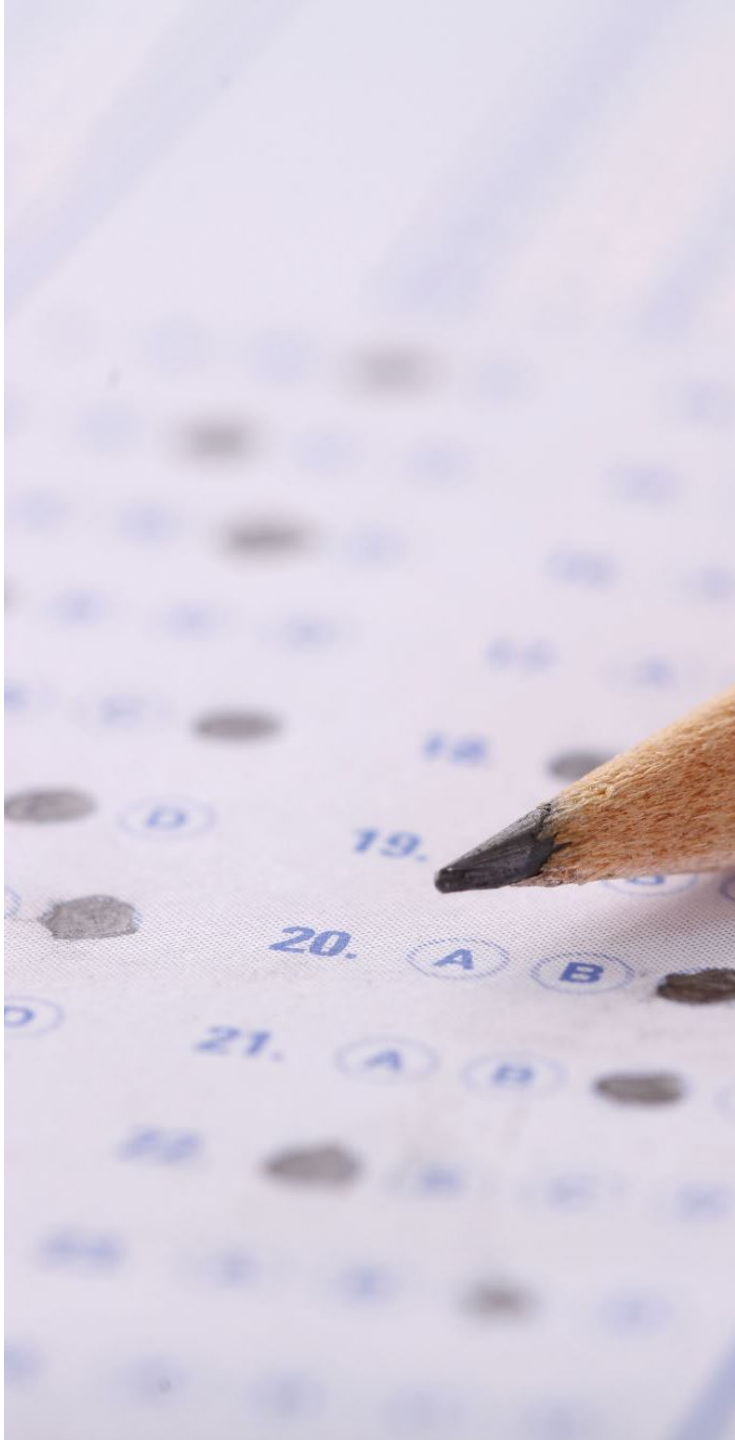
  1: 35% of problems found

  3-5: 65-75% of problems found

  Diminishing returns thereafter

QUERY: SURVEY

- Time consuming to produce and must be done correctly
- Once produced - can provide a vast body of information
- Good for attitude measurement across large groups
- Good design:
  - Mixture of open & closed questions
  - Structure questions carefully: general to detailed
  - Not too long
  - Try it out prior to its actual use
- Problems: loaded terms; suggested responses; lack of precision; subject embarrassment



QUERY: INTERVIEW

Interviews:

- More active user involvement
- More exploration, richer data
- Self report data
- Structured / Unstructured
- More time consuming to conduct
- Evaluator experience / bias
- More difficult to analyse

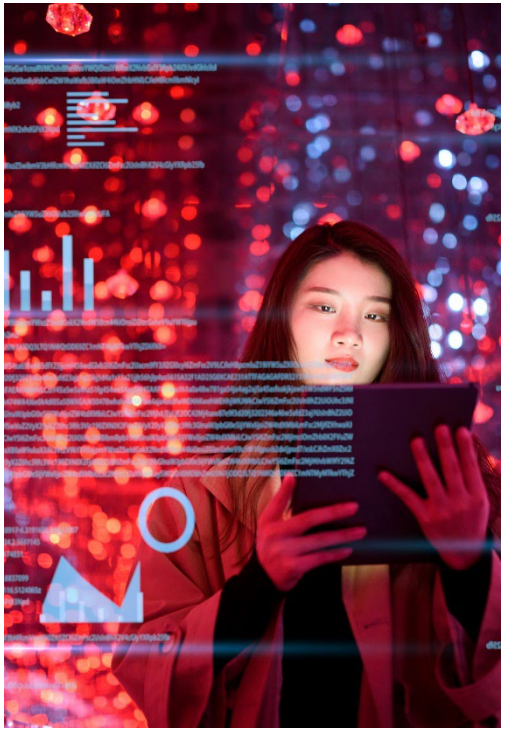
SELF-REPORT SOFTWARE INSTRUMENT: PREMO

- PrEmo is a non-verbal self-report software instrument that measures 14 emotions that are often elicited by product design
- Emotional responses difficult to measure because their nature is subtle (low intensity) often mixed (i.e. more than one emotional response at the same time)
- Does not rely on words
- Each of the emotions is portrayed by an animation of dynamic facial, bodily, and vocal expressions
- For use in internet surveys, formal interviews, and qualitative interviews, e.g., to identify the concept with the most pleasant emotional impact as a discussion tool in consumer interviews

EVALUATION ANALYTICS: PREMO

KAREN YOUNG, SCHOOL OF COMPUTER SCIENCE, UNIVERSITY OF GALWAY, 2023-2024

ANALYTICS



- **ESM**: Experience Sampling Method
- A set of methods designed to repeatedly request people to document and report their thoughts, feelings and actions outside the laboratory, within the context of their everyday life
- Sending requests can be timed, random, triggered by context variables or triggered by user actions
- Requests and answers can be sent through a variety of devices (internet, web browser, text messaging on mobile phone, etc.)
- Strengths: Does not depend on recall (no memory effects)
- Provides insights into variability over time
- Weakness: Requires high compliance, high effort (interferes with activity)

OBSERVATIONAL STUDIES

- Important not to disturb the way an individual works on a task - watch extraneous conditions: Hawthorne effect
  - Clear Goals: scope, hypothesis?
  - Ethics: consent, stress, time limit
  - Pilot your study: ideally two pilots: colleague, one real user
  - Collect your results: co-operative evaluation, thinking aloud?
  - Debrief: holistic review
- Examples?

THINKING-ALOUD (MONK ET AL)

- One popular and successful method is the ***Thinking-Aloud*** method (Monk et al)
- Really powerful insights
- Users perform tasks - you monitor/observe their performance and thought processes while doing tasks
- Need to prompt people: not natural
- Decide ahead of time what you will and won't help on
- Try to avoid specific questions
- Not useful for collecting numeric task completion data: delays user in task
- Also can use usability labs - with outside observers

CO-OPERATIVE EVALUATION

- A procedure that encourages collaboration between users and designers: users ask evaluator questions and evaluator questions them about their understanding, thought processes etc.
- Natural process: work through a set of tasks
- Identify most important improvements to consider
- Steps:
  - Recruit users
  - Prepare tasks
  - Interact and record
  - Debrief users
  - Summarise your observations

USABILITY TESTING

- Goals & questions focus on how well users perform tasks with the product
- Comparison of products or prototypes is common
- Focus is on time to complete task & number & type of errors
- Data collected by video & interaction logging
- User satisfaction questionnaires & interviews provide data about users' opinions

USABILITY PRINCIPLES

- Similar to design principles, except *more prescriptive*
- Used mainly as the basis for evaluating systems
- Example principles:
  - Visibility of system status
  - Match between system and the real world
  - Consistency and standards
  - Error prevention
  - Recognition rather than recall
  - Flexibility and efficiency of use

USABILITY?

"However, the reality is that we all continue to experience frustrations when using interactive digital technologies, and often we would say that we do find them difficult to use. ***Even so, frustrating user experiences may not be due to some single abstract construct called 'usability', but instead be the result of unique complex interactions between people, technology and usage contexts.*** Interacting factors here must be considered together. It is not possible to form judgements on the severity of isolated usage difficulties, user discomfort or dissatisfaction. Overall judgements on the quality of interactive software must balance what can be achieved through using it against the costs of this use. There are no successful digital technologies without what could be usability flaws to some HCI experts (I can *always* find some!). Some technologies appear to have severe flaws, and are yet highly successful for many users. Understanding why this is the case provides insights that move us away from a primary focus on usability in interaction design." (Cockton)

USABILITY ENGINEERING

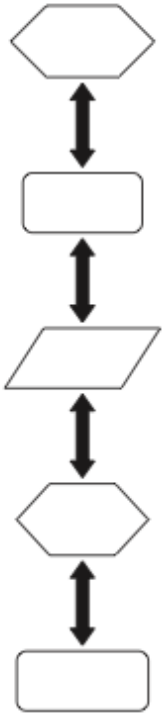
- All design decisions should be conscious and visible
- Need ways to measure results against agreed criteria
- Requires the adoption of a good list of attributes that are measurable
- Usability specification will state how criteria will be measured, what they are, what the pre-conditions are
- Should also specify worst case, lowest acceptable level, planned case, best case and "now" level

USABILITY CHECKLIST

Usability checklist (learnability, throughput, satisfaction):

- Time taken to complete the task
- Percentage of task completed
- Ratio of success to failure
- Time spent dealing with errors
- Use of help and on-line documentation: frequency
- Percentage of favourable/unfavourable user comments
- Number of repetitions or failed commands
- Number of commands not used
- Number of good features recalled by user

USABILITY & FIELD STUDY



Field study to evaluate initial design ideas and get early feedback

Make some design changes

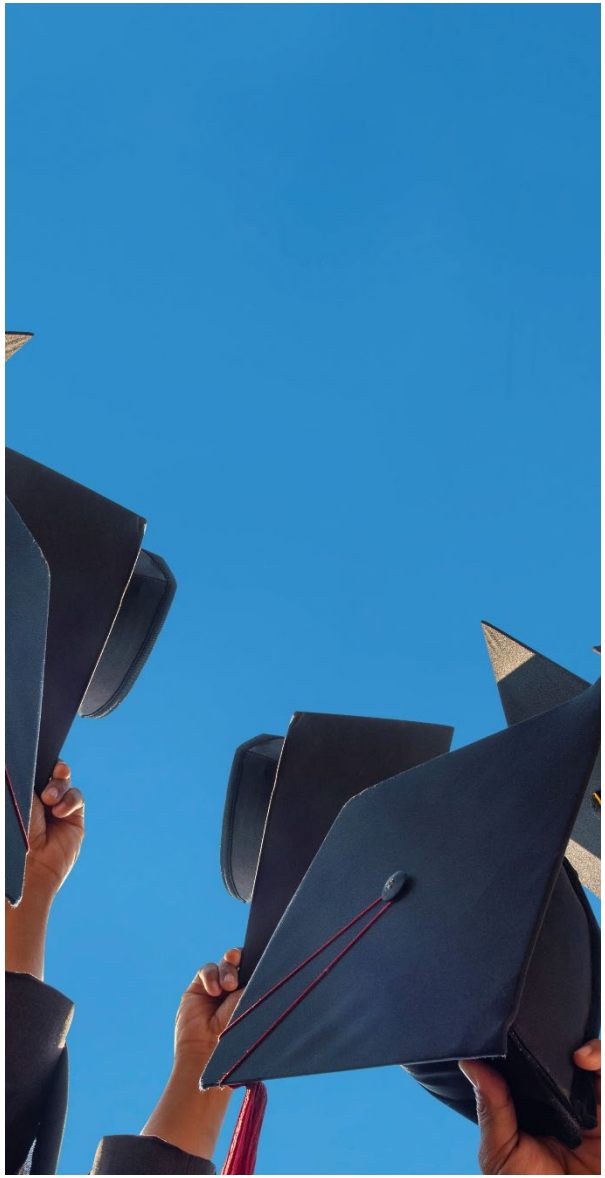
Usability test to check specific design features

Field study to see what happens when used in natural environment

Make some final design changes

USABILITY: USERS?

- Graduating Users: Beginners to Intermediate to Expert
- "*Information in the world and information in the head*"; Donald Norman, *DOET*
- ***World vectors*** are required by beginners and more expert users for advanced or seldom-used functions
- ***Head vectors*** are used extensively by intermediates and even moreso by experts
- New users happy with world vectors, but as they progress they develop working sets: provide a head vector as well as world vector, and a path by which user can learn head vector

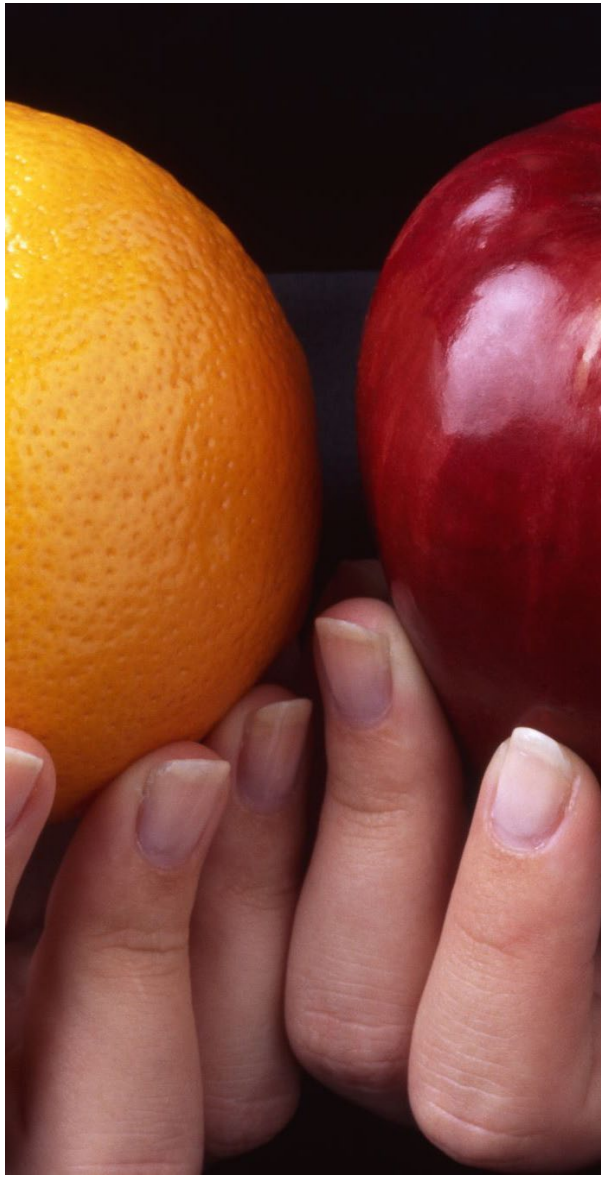


EVALUATION: EXPERIMENTS

- Controlled evaluation of specific aspects of interactive behaviour
- Hypothesis: chosen by evaluator
- Subjects?
- Variables: Independent (controlled by evaluator), Dependent
- Experiment Design: Comparative/ Absolute
  - Identify the problem and formulate hypothesis
  - Design & execute experiment
  - Examine data from experiment
  - Communicate the results

COMPARATIVE EVALUATION

- Is Interface X better than Interface Y?
- What is ***better***?
- Answer: "it depends"
- What does it depend on?
- Dependent Variables: Measures:
  - Time
  - Errors
  - Recall
  - Conversions (purchases)
  - Emotional response



EVALUATION: EXPERIMENTS

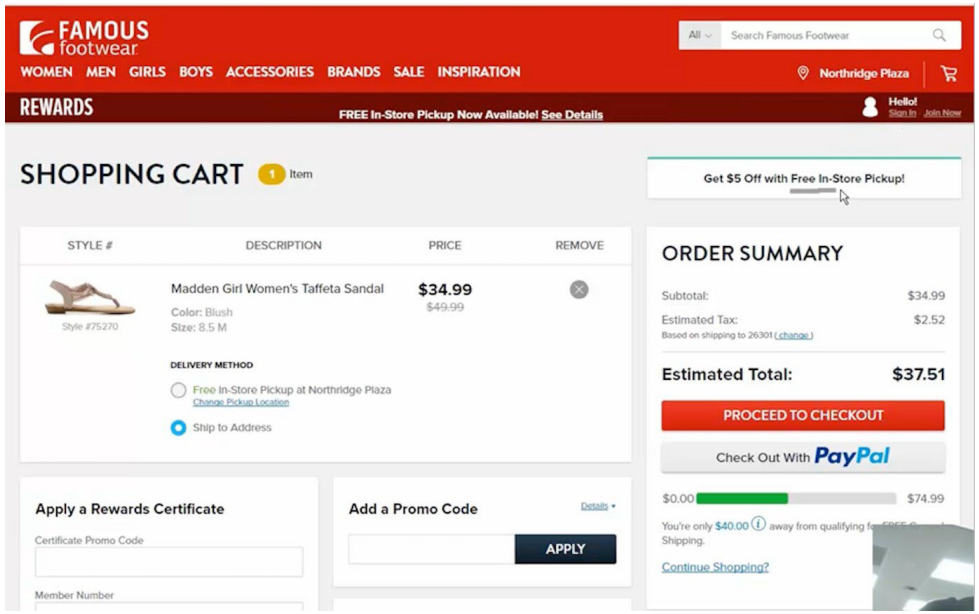
- **Manipulations** (independent variables – controlled by experimenter; e.g. colour, size)
- **Measures** (dependent variables; e.g. duration, errors, feelings)
- Problems which can interfere with results:
  - Practice effect
  - Fatigue effect
  - Order effect

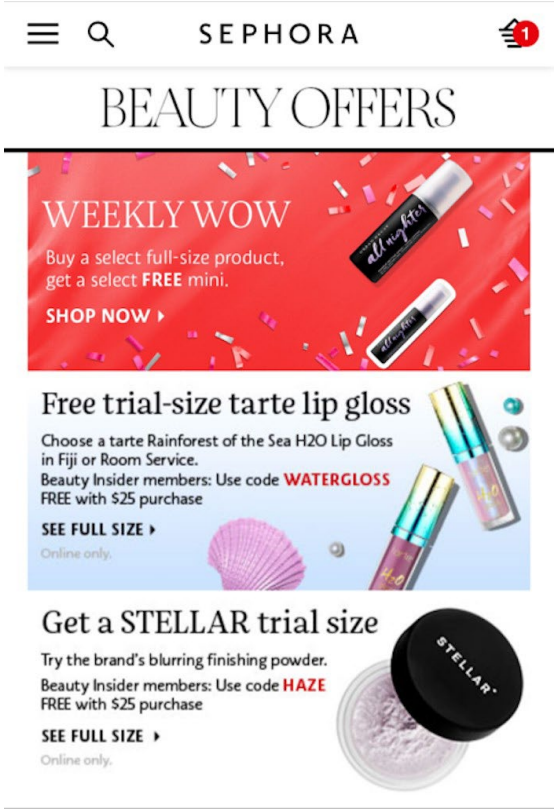
EVALUATION: EXPERIMENTS

- An example: evaluating icon design
- Hypothesis : User will remember the natural icons more easily than the abstract ones
- Null hypothesis: no difference between recall of the icon types
- Between-subjects or Within-subject design? Why?

ONLINE EXPERIMENTS

- Online Experiments: randomly split traffic to website between 2 / more UI versions
- What do you want to measure: meaningful statistics: click throughs, conversions, etc.
- Key Findings:
  - **Commitment Escalation**: ask them to commit a little upfront, then add a little more later
  - Small, insignificant changes: big impact (coupons, company name etc.)
  - Our **expectations** are often wrong: e.g. images preferred over video


FAMOUS footwear
WOMEN MEN GIRLS BOYS ACCESSORIES BRANDS SALE INSPIRATION
Search Famous Footwear
Northridge Plaza
REWARDS
FREE In-Store Pickup Now Available! See Details
Hello! Sign In Join Now
SHOPPING CART 1 Item
Get $5 Off with Free In-Store Pickup!
STYLE # DESCRIPTION PRICE REMOVE
Madden Girl Women's Taffeta Sandal
Style #75270
Color: Blush
Size: 8.5 M
$34.99
$49.99
DELIVERY METHOD
Free In-Store Pickup at Northridge Plaza
Change Pickup Location
Ship to Address
ORDER SUMMARY
Subtotal: $34.99
Estimated Tax: $2.52
Based on shipping to 26301 (change)
Estimated Total: $37.51
PROCEED TO CHECKOUT
Check Out With PayPal
$0.00 $74.99
You're only $40.00 away from qualifying for FREE Shipping.
Continue Shopping?
Apply a Rewards Certificate
Certificate Promo Code
Member Number
Add a Promo Code
Details
APPLY


SEPHORA
BEAUTY OFFERS
WEEKLY WOW
Buy a select full-size product, get a select FREE mini.
SHOP NOW
Free trial-size tarte lip gloss
Choose a tarte Rainforest of the Sea H2O Lip Gloss in Fiji or Room Service.
Beauty Insider members: Use code WATERGLOSS
FREE with $25 purchase
SEE FULL SIZE
Online only.
Get a STELLAR trial size
Try the brand's blurring finishing powder.
Beauty Insider members: Use code HAZE
FREE with $25 purchase
SEE FULL SIZE
Online only.
STELLAR

WEB EVALUATIONS

- Web site evaluations: ready-to-hand measures of web server logs – easy but appropriate?
- What is easy to measure via a web server is rarely what is needed for meaningful relevant user experience evaluation
- **PULSE** measures are Page views, Uptime, Latency, Seven-day active users (i.e. count of unique users who used system at least once in last week), and Earnings.
- Research at Google (Rodden at al, 2010): **HEART** UX measures
- **HEART**: Happiness, Engagement, Adoption, Retention, and Task success

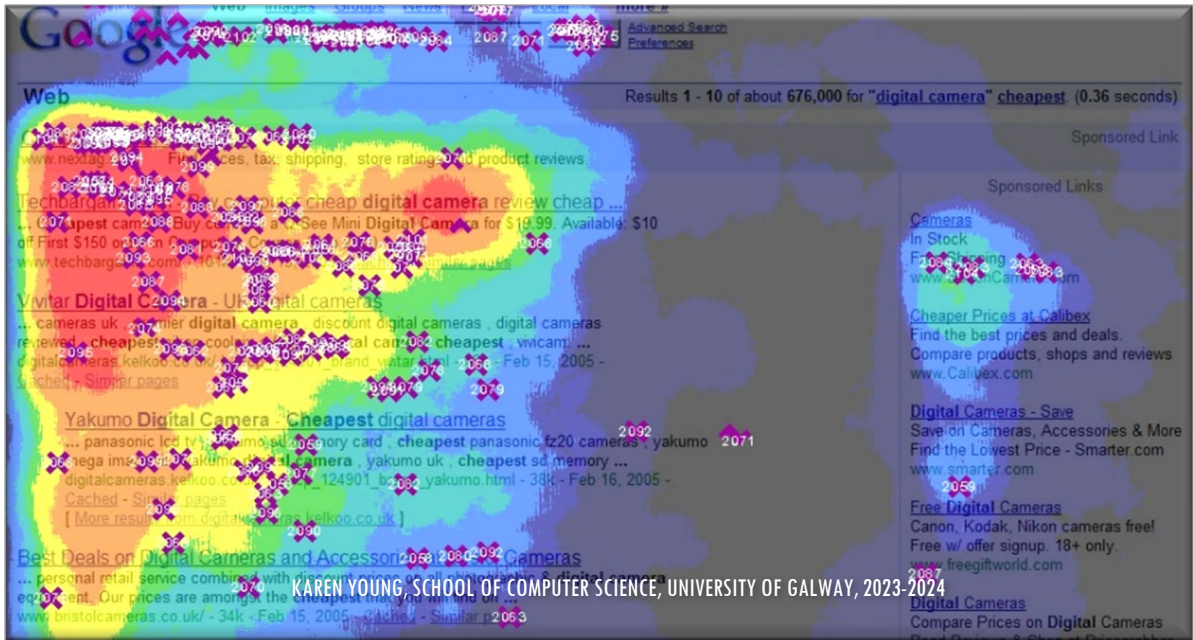
PHYSIOLOGICAL MEASURES

- Physiological measurements: Why?
- Heart activity; blood pressure, volume and pulse
- Activity of the sweat glands; galvanic skin response(GSR)
- Electrical activity in muscle; electromyogram (EMG)
- Electrical activity in the brain; electroencephalogram(EEG)

EYE TRACKING

- High tech method; requires equipment
- Strengths: hard-to-articulate behaviours; compelling visualising data; exciting clients
- Molich (2008: CUE 7 Study): eye tracking did not identify any new issues over inexpensive methods
- Eye tracking: fixations don't communicate meaning: interpretation (where, not why)
- Problems: time, cost, complexity, technical
- Used effectively, can provide insights
- Not essential to usability testing; Decision made according to testing goals and considerations

EYETRACKING



MOBILE EVALUATION

- Mobile usage:
  - Commuting
  - At home
- Drivers for engagement:
  - Daydreaming
  - Quick wins, help me now
  - Monitoring and instant gratification
- Barriers:
  - Security
  - Screen size
  - Connectivity
- Most effective: short, interrupted interactions that can be integrated into routine; time saving or entertainment

RESULTS

EVALUATION: RESULTS

- How to capture evaluation results?
  - What type of data: quantitative (less time consuming), qualitative
  - Note taking: pen & paper, computer
  - Video recording: rich data, can grab success / challenge and share with others, time consuming
  - Screen recording: good for capturing UI, not users' facial expressions
- Debrief users afterwards

EVALUATION: RESULTS

- Observation & interviews:
  - Notes, pictures, recordings
  - Video
  - Logging
- Analyses
  - Categorized
  - Categories can be provided by theory
    - Grounded theory
    - Activity theory

EVALUATION: RESULTS

- What to do with collected evaluation data?
  - What does my data look like?
  - What are overall numbers?
  - Are the differences real: Pearson chi-squared test (discrete rate data), t-tests (continuous), anova tests (> 2 conditions): which vacuum cleaner, which input device is faster for input?
- Use multiple evaluation methods together: combination better than sum of parts

EVALUATION: CONSIDERATIONS

- **Validity**: does the method measure what it is intended to measure?
- **Reliability**: does the method produce the same results on separate occasions?
- **Ecological validity**: does the environment of the evaluation distort the results?
- **Biases**: Are there biases that distort the results?
- **Scope**: How generalizable are the results?

EVALUATION: CHALLENGES

- Evaluation Method
  - Suitability of Method: essential vs. relational
  - Test Environment
  - Completeness of prototype
- Users
  - Previous experience
  - Motivation and interest
  - Sample size
- Evaluators
  - How experience?
  - Are they biased?

DECIDE FRAMEWORK

- Well-planned evaluations are driven by clear goals and appropriate questions
- DECIDE Framework (Preece, Rogers & Sharp):
  - **D**etermine the overall goals the evaluation addresses
  - **E**xplore the specific questions to be answered
  - **C**hoose the evaluation paradigm and techniques to answer the questions
  - **I**dentify the practical issues: e.g. selecting participants, finding evaluators, equipment etc.
  - **D**ecide how to deal with the ethical issues
  - **E**valuate, interpret, and present the data

HUMAN COMPUTER INTERACTION

Lecture 10 Review: Evaluation

- Challenges of Evaluation
- Evaluation: Methods & Measures
  - Heuristic
  - Usability
  - Experiments
- Evaluation Results
- Evaluation Framework: DECIDE

EVALUATION

KAREN YOUNG, SCHOOL OF COMPUTER SCIENCE, UNIVERSITY OF GALWAY, 2023-2024